Localizing Ontologies in OWL

Wim Peters[1], Elena Montiel-Ponsoda[2], Guadalupe Aguado de Cea[2], and Asunción Gómez-Pérez[2]

[1]University of Sheffield, U.K.
W.Peters@dcs.shef.ac.uk
[2]Universidad Politécnica de Madrid (UPM)
emontiel@delicias.dia.fi.upm.es
{lupe, asun}@fi.upm.es

Abstract. This paper presents a model for linguistic/terminological information, which can be used in tandem with an ontological model, in order to link lexicalizations and concepts. The main aim of the proposed model is to provide multilingual information to ontologies. Interoperability with existing standard models of terminological description as well as access to authoritative linguistic resources are crucial aspects that have been considered in the design of the proposed model.

Keywords: ontology localization**,** standardization, resource interoperability, ontology-lexicon interface, multilingual ontologies

1 Introduction

Since the beginning of the ontological engineering in the last decade of the 20th century, ontologies have rapidly widened its field of application and are nowadays considered one of the main pillars in the construction of the Semantic Web. Ontologies have proved to be the most reliable resources to represent agreed domain knowledge, enabling better communication and performance in semantic web applications.

However, the huge increase in the development of ontologies (see e.g. the DAML ontology library[1]) and their pervasive use in a wide variety of domains have shown the need for addressing the issue of the provision of ontologies with linguistic data for determining possible lexicalizations for concepts and addressing multilinguality. This link between lexical and ontological knowledge is in many cases the only way to make conceptual knowledge shareable across ontologies and humans [7]. Since many ontologies do not contain definitions, but rely, for the purpose of determining their ontological nature, on labels for concepts, and their conceptual context within the ontologies they occur in, a great burden is put on the interpretation of the lexicalizations.

[1] http://www.daml.org/ontologies/

Ontologies themselves are conceptual constructs without linguistics. From a formal ontological point of view, concepts are abstract notions whose labels are arbitrary. The lexical senses of the lexicalizations that function as labels for these concepts, are only considered to be evocative or indicative of the ontological meaning of the concepts. There is an implicit mapping assumption between lexical and conceptual knowledge, which underlies "ontology lexicalization", namely that (intensional) senses from a lexical model are mapped to (extensional) interpretations on ontology elements (individuals, classes, restrictions, properties). The lexical semantic content of the lexicalizations, originating from linguistic/terminological resources such as term banks, thesauri and dictionaries, is considered to be lightweight, and in need of formalization.

In order to capture and represent the interplay between conceptual and lexical meaning, we need to define a model which links both types of meaning by means of an ontological module on the one hand, and a linguistic/terminological module on the other.

2 The Ontological Meta-model

For the representation of conceptual knowledge, we adopt the OWL ontology meta-model[2] [2], which is illustrated in figure 1.

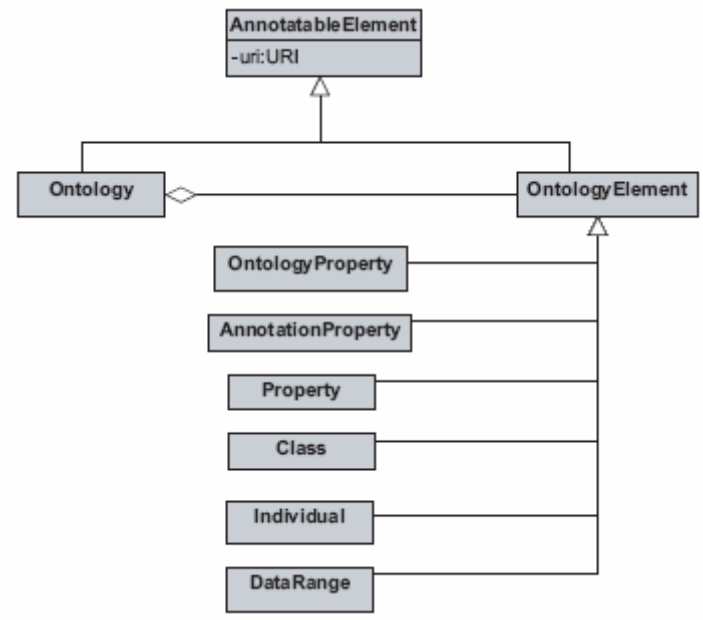

Fig. 1. Main elements of the OWL Ontology Meta-model

[2] http://owlodm.ontoware.org/OWL1.1

This figure shows the central part of the OWL meta-model, which follows the Description Logic (DL) paradigm. `OntologyProperty`, `AnnotationProperty`, `Property`, `Class`, `Individual` and `DataRange` are all Ontology Elements. It has been our aim to design a linguistic model that allows the association of linguistic and terminological data with each `OntologyElement`.

3 Requirements for the Linguistic Model

Building a model that provides linguistic and terminological information to ontology elements needs to take the following important issues into account: accessibility, resource interoperability and multilinguality.

3.1 Accessibility

The model should enable the user to browse resources, access the linguistic information contained in them, and select relevant lexicalizations for definition and formation of concepts.

An example of how resources can be used to suggest conceptualizations and lexicalizations in ontology design is provided by tools such as LabelTranslator [4] and OntoLing [8] Such tools support the addition of linguistic -or multilingual- data to already existing ontologies or ontologies under development. The main idea behind both systems is to facilitate access to external semantic resources for the domain expert, translator or terminologist, and offer a "semi-automatic extension" or enrichment of ontologies with linguistic data.

The current version of the LabelTranslator platform allows access to authoritative multilingual resources such as EuroWordNet[3] [10] and returns possible lexicalizations, translations and definitions. Then, the user selects the most suitable linguistic information for the ontology element in question. This tool is currently being extended within the NeOn[4] project in order to increase the typology and number of accessed resources, and, most importantly, to improve the results by introducing disambiguation and translation algorithms that pursue an automation of the process.

OntoLing implements a navigation system to e.g. wordnet databases, and allows the choice of individual lexical items or entire branches, and their promotion to ontology classes or subsumption hierarchies. Alternatively, lexical items can be used to label existing classes. OntoLing implements an "on-the-fly re-engineering method", which is a good example of exploiting lexical resources without caring too much about the general reengineering design, which is left to user choices, when a specific design need arises. The Linguistic Watermark[5] package contains the underlying model, which functions as the standard representation of the resources contained in the package, and which is suggested as the standard reference model for inclusion of additional resources by the user.

[3] http://www.illc.uva.nl/EuroWordNet/
[4] http://www.neon-project.org/web-content/
[5] http://ai-nlp.info.uniroma2.it/software/LinguisticWatermark/index.html

Both models cover a number of linguistic/terminological information units, but are by no means exhaustive, and cannot cover all information units needed by ontology engineers who want to access resources on the basis of widely used standards for linguistic/terminological information exchange (see next point).

3.2 Resource Interoperability

Lexical knowledge should be encoded following standard models in order to guarantee interoperability with existing and proposed standards for the representation and integration of terminological and linguistic knowledge.

For this purpose, many standardization initiatives have been developed in order to capture terminological and linguistic information that can be re-used for various purposes. As the most important initiatives we mention a number of standards from the International Organization for Standardization (ISO)[6], which capture terminological and linguistic information, and need to be taken into account:

The TMF framework[7] (and the associated TermBase eXchange format; TBX[8]) captures the underlying structure and representation of computerized terminologies. Both formats make use of ISO 12620 data categories[9].

The Lexical Markup Framework (LMF; ISO/CD 24613) [5] is an abstract meta-model that provides a common, standardized framework for the construction of computational lexicons. The LMF ensures the encoding of linguistic information in a way that enables reusability in different applications and for different tasks. The LMF provides a common, shared representation of lexical objects, including morphological, syntactic, and semantic aspects. It is under development and expected to be declared a standard in 2008.

Within the ISO, the Technical Committee 37/SC 4 is in charge of the "Language resource management", and Work Group 3 (WG 3) of this committee is currently dealing with the "Multilingual information representation". For this purpose, the WG 3 has already proposed a standard called MLIF (Multi Lingual Information Framework) [11] with the aim of providing a common platform for integrating the above mentioned standards. In this sense, MLIF could be considered a "meta-standard" that allows for the interaction of different representation models, in which the designer can select which models to use depending on the linguistic needs of the end resource.

SKOS Core[10] (Simple Knowledge Organization Systems) has been developed within the W3C framework, and provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, 'folksonomies', other types of controlled vocabulary, and also concept schemes embedded in glossaries and terminologies.

There have been many other standardization initiatives, such as the Text Encoding Initiative[11], which does not make detailed proposals for lexical tag sets, but describes

[6] www.iso.org
[7] http://www.loria.fr/projets/TMF/
[8] http://www.lisa.org/standards/tbx/
[9] http://www.ttt.org/oscar/xlt/webtutorial/
[10] http://www.w3.org/2004/02/skos/core/
[11] http://www.tei-c.org.uk/

the structure of a dictionary entry in detail. Various standardization efforts such as EAGLES[12] and ISLE[13] worked out concrete proposals for standard lexical structures, according to which the PAROLE[14] multilingual lexicons have been encoded in a uniform way for 12 languages.

Within the area of terminology and machine translation, other relevant work was undertaken by OLIF (Open Lexicon Interchange Format)[15]. It defines a large number of lexical features, but does not make statements about their structural embedding.

In order to obtain interoperability between all these representational standards, we need to ensure a maximum level of overlap between the information units for linguistic and terminological description defined in each standard. For the linguistic model we present in this paper, we have analyzed the standards mentioned above, and mainly based ourselves on TMF and TBX, the standards with the greatest impact in the terminological field, and LMF, which integrates previous initiatives in the linguistic arena.

3.3 Multilinguality

Recently, the need for providing multilinguality to ontologies has emerged as one of the main priorities in the Knowledge Engineering research. The incremental use of knowledge based systems has raised the need of expressing knowledge in a way that can be understood by people coming from different cultures and speaking different languages, i.e, the need for having to adapt knowledge for specific cultural and language universes. The process of adapting an ontology to a concrete language and culture community has received the name of "ontology localization".

The coverage of multilinguality is an increasingly important issue. The ontology library OntoSelect[16] reports the existence of 36 multilingual ontologies out of the total amount of 1420 ontologies that it contains (2.5%). Most of these ontologies containing multilingual labels lack consistency in their coverage of languages, which are not the original language of the resource (English in most cases). The number of multilingual ontologies is expected to grow fast in the coming years, and the growing interest in multilinguality as a challenge for knowledge based approaches manifests itself in different ways, from multilingual information retrieval, query answering systems and machine translation [1] [9].

This has been the main motivation for our research, in which we have tried to design a model that captures linguistic data in such a way that it permits, on the one hand, to maximize the correspondence between ontological conceptualization and linguistic/terminological standardization, and, on the other hand, to enrich the ontology with natural language information in order to localize the ontology and make it suitable for a specific culture and language community.

---

[12] http://www.ilc.cnr.it/EAGLES96/edintro/edintro.html
[13] http://www.mpi.nl/ISLE
[14] http://www.elda.org/catalogue/en/text/doc/parole.html
[15] http://www.olif.net
[16] http://olp.dfki.de/OntoSelect/

4 The Linguistic Information Repository (LIR)

The linguistic/terminological meta-model in Figure 2 has been designed from the perspective of the ontology engineer. It takes relevant linguistic and terminological knowledge from resources into account, such as term banks, thesauri and dictionaries, in order to create a linguistically/terminologically informed link between intra- and extra-ontological information.

It is a structured, eclectic set of linguistic and terminological data categories, built up on the basis of existing standards. This ensures interoperability with these standards, and a maximum level of acceptance within the user communities, active in the combined fields of linguistics, terminology and ontology engineering.

It is extensible in the sense that it will be able to accommodate any additional data categories deemed useful for an ontology engineer editing lexicalizations and browsing available linguistic information such as alternative lexicalizations and translations. For instance, the class `UsageContext` (see figure 2 below) can be extended with new subclasses from the TBX data category proposal[17], such as definitional and associative context. Also, further morphological and syntactic decomposition such as headword identification and stemming can be included [3]. Moreover, foreseeable future developments, such as a typology of definitional structure, can be added without the stamp of official standardization, while still building on standard information structures.

The model contains the following classes:

1. `LexicalEntry`: a lexeme, which is a unit of form and meaning[18].
2. `Sense`: a language-specific unit of intensional lexical semantic description.

The addition of the attribute xml:lang to `Sense` allows us to model language specific meaning.

3. `PartOfSpeech`: The grammatical class of the `LexicalEntry`
4. `Lexicalization`: a word form. This class corresponds with the LMF class Form Representation. It means that the lexicalizations of concepts are deemed word forms rather than lemmas or citation forms, and are therefore allowed to be inflected forms, such as plurals.

`Lexicalization` has a number of attributes and relations selected from TMF[19] and TBX-Lite[20], and split up into a set of Boolean attributes and a number of relations between Lexicalization classes.

5. `Definition`: terminological/linguistic sense description.
6. `Source`: the provenance of the linguistic/terminological information.
7. `UsageContext`: example usages of the lexicalization in texts.
8. `Note`: any noteworthy information in free text form.

[17] http://www.lisa.org/standards/tbx/
[18] See e.g. http://en.wikipedia.org/wiki/Lexeme
[19] http://www.ttt.org/oscar/xlt/webtutorial/datcats02.htm
[20] http://www.lisa.org/standards/tbx/

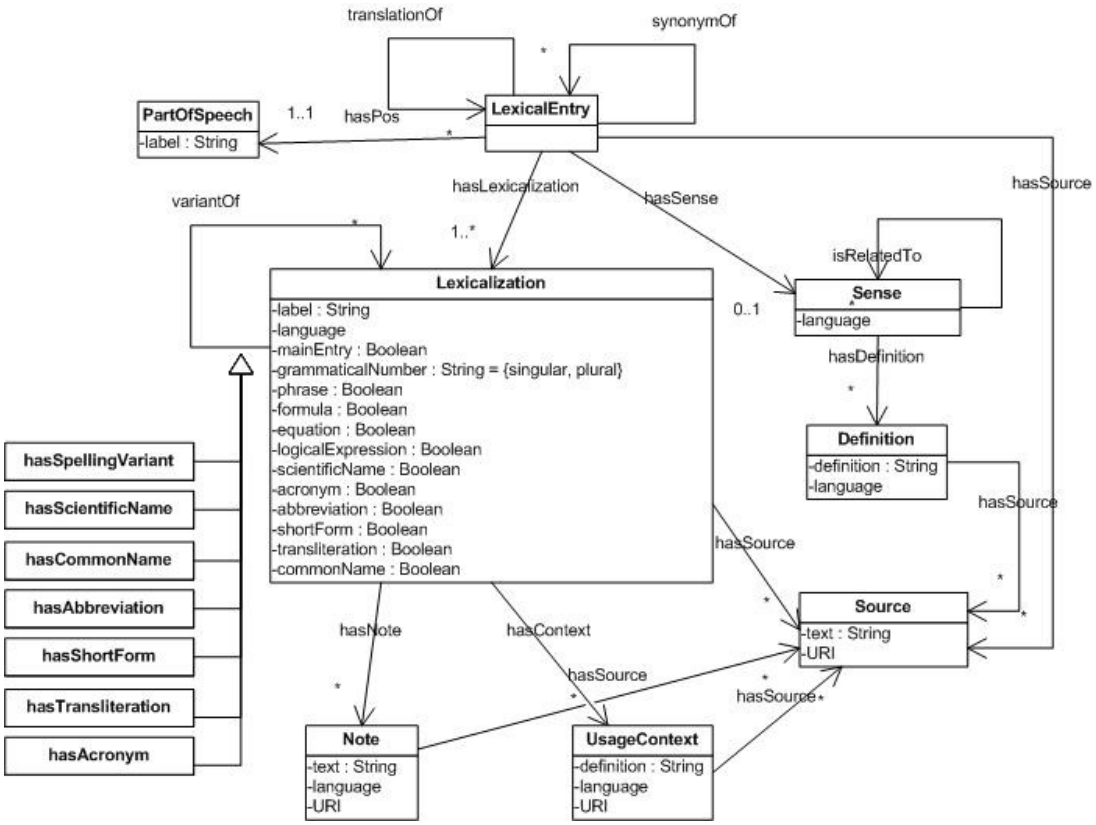

Fig. 2. Linguistic Information Repository (LIR) model

5 Linking Ontological and Linguistic Knowledge

As mentioned above, linking conceptual and linguistic knowledge involves two separate models for these two types of knowledge: the ontological meta-model and the linguistic model. Classes, properties or individuals of the ontological meta-model can be provided with lexicalizations from the separate linguistic model in the form of lexemes, i.e. units of form and meaning. This model contains a set of data categories that captures all the relevant linguistic/terminological information associated with concepts such as lexicalizations, lexicalization types and multilinguality.

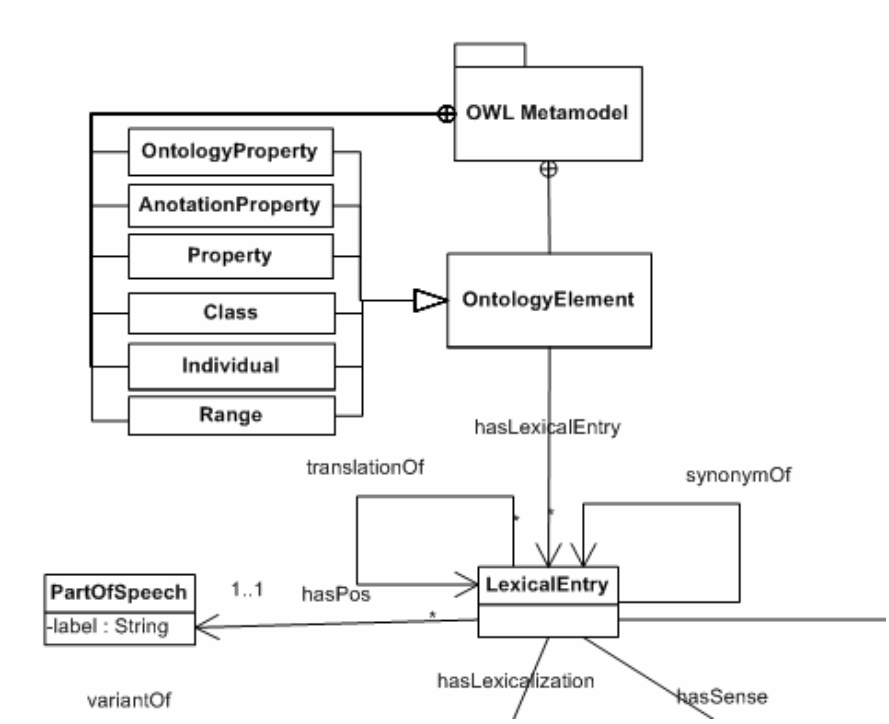

Fig. 3: The Link between Ontological and Lexical Knowledge

Each ontology element of the ontology meta-model is linked to one or more lexicalizations from one or multiple languages. This is illustrated in Figure 3. The link consists of a general relation `hasLexicalEntry`, with, as yet, no semantic characterization apart from "is lexicalized by". The ontology engineer decides if a lexical entry applies to a concept to a sufficient level of satisfaction. If we consider, for instance, the use of semantic and conceptual features in the description of concept and sense, it is possible to create a further sub-classification of this correspondence relation along the lines of set relations such as subset, overlap, and even disjointness [6].

6 Discussion and Future Work

The model we have presented in this paper regulates the interplay between ontology and linguistic/terminological information within a semantic web setting, by using established and new standards from the linguistic and terminological field. It offers the possibility to further define the nature of the link between lexical semantic and conceptual knowledge representation. This link can range from full equivalence for light weight ontologies to mere complementation in the case of a general language resource to a highly specialized terminological field.

The model will be available in OWL format, and is therefore extensible. Additional linguistic typology can be accommodated by means of the integration of new classes into this ontology. This will be based on the requirements of the ontology editor. For instance, if a greater morphological analysis of ontology labels is required, it may be decided to integrate the appropriate LMF module.

The model lays the foundations for further development, in that it allows a text-based characterization of its classes by means of Note. This may serve as the basis for future formalization of more fine-grained semantic distinctions in e.g. translational and conceptual equivalence. Also, it allows a gradually emerging typology of the semantically underspecified `haslexicalEntry` link between `OntologyElement` and `LexicalEntry`.

Our model is more top-down (standardized) resource based approach to linguistic modeling than LabelTranslator's and Ontoling's bottom-up approach. Data structures are pre-defined, which means that resource-specific information units from other, widely used, de facto standard representations, such as TEI and JAVADICT, need to be linked up by associating their units of description with LIR data categories. The addition of information from other resource-specific formats will need to follow the same route. The advantage of adopting a standard-based adoption of linguistic data categories as opposed to e.g. JAVADICT is that the standard data categories offer a suitable breadth and depth of coverage of linguistic phenomena. Using these as entry points for ensuring interoperability between resources provides a more homogeneous and theoretically more widely agreed upon coverage, as opposed to an organically growing and mutually enriching set of inter-linked resource-specific representations, which cover only part of the linguistic spectrum.

The model has been checked against the requirements from a variety of resource formats, in particular EuroWordNet, the TBX/TMF specifications, and the LMF and SKOS architectures.

It is foreseen that the next version of LabelTranslator will be compliant with this model, and therefore the power to capture existing and future resource-specific, non-standard authoritative and standard structures.

Acknowledgments. This work has been funded by the NeOn project (Life Cycle Support for Ontologies; FP6-027595. We would like to thank Margherita Sini and Aldo Gangemi, José Ángel Ramos Gargantilla and Mari Carmen Suárez-Figueroa for their feedback.

References

1. Benjamins, V.R., Contreras, J., Corcho, O., and Gómez-Pérez, A., SixChallenges for the Semantic Web, SIGSEMIS bulletin, April (2004)

2. Brockmans, S., Haase, P. and Studer, R., A MOF-based Metamodel and UML Syntax for Networked Ontologies. Intl. Semantic Web Conf. Georgia, US (2006)

3. Buitelaar, P. Sintek, M., Kiesel, M.: A Lexicon Model for Multilingual/Multimedia Ontologies In: Proceedings of the 3rd European Semantic Web Conference (ESWC06), Budva, Montenegro (2006)

4. Declerck, T., Vela, O., Gómez Pérez, A., Gantner, Z., Manzano Macho, D.: LabelTranslator: Multilingualism in Ontologies, in Poster/Demo Track of the Fourth International Semantic Web Conference (ISWC 2005), Galway, Ireland (2005)

5. Francopoulo, G, George, M. Calzolari, N. Monachini, M. Bel, N. Pet, M. Soria, C.: Lexical Markup Framework (LMF) In: Proc. of the International Conference on Language Resources and Evaluation (LREC) , Genoa, Italy (2006)

6. Holi, M. and Hyvönen, E.: "Probabilistic information retrieval based on conceptual overlap in semantic web ontologies". In Proc. Finnish Artificial Intelligence Conference (FAIS'04), vol. 2, Finnland (2004)

7. Pazienza, M.T., Stellato, A.: Exploiting Linguistic Resources for building linguistically motivated ontologies in the Semantic Web. Second Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006), held jointly with LREC2006 ,Magazzini del Cotone Conference Center, Genoa, Italy (2006)

8. Pazienza, M.T., Stellato, A.: Linguistic Enrichment of Ontologies: a methodological framework. Second Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006), held jointly with LREC2006, Genoa, Italy (2006)

9. Peñas, A., Gonzalo, J.(eds.): Acceso a información multilingüe. Número monográfico de la Revista Iberoamericana de Inteligencia Artificial, Vol. 8. nº 22. (2004)

10. Vossen, P., Peters, W., Díez-Orzas, P.: The Multilingual design of the EuroWordNet Database, In: Kavi Mahesh (ed.) Ontologies and multilingual NLP, Proceedings of workshop at IJCAI-97, Nagoya, Japan (1997)

11. Cruz-Lara, S., Bellalem, N., Ducret, J., Kramer, I.: Standardizing the Management and the Representation of Multilingual Data: the MultiLingual Information Framework. LR4TransIII, Genoa, Italy (2006)